Evaluating Legal Argument Instruction with Graphical Representations using LARGO

Niels PINKWART[1], Vincent ALEVEN[2], Kevin ASHLEY[3], and Collin LYNCH[4]
[1]*Clausthal University of Technology, Computer Science Institute, Germany*
[2]*Carnegie Mellon University, HCI Institute, Pittsburgh PA, USA*
[3]*University of Pittsburgh, LRDC and School of Law, Pittsburgh PA, USA*
[4]*University of Pittsburgh, Intelligent Systems Program, Pittsburgh PA, USA*

Abstract. Previous research has highlighted the advantages of graphical argument representations. A number of tutoring systems have been built that support students in rendering arguments graphically, as they learn argumentation skills. The relative tutoring benefits of graphical argument representations have not been reliably shown, however. In this paper we present an evaluation of the LARGO system which enables law students graphically to represent examples of legal interpretation with hypotheticals they observe while reading texts of U.S. Supreme Court oral arguments. We hypothesized that LARGO's graphical representations and advice would help students to identify important elements of the arguments (i.e., proposed hypotheses, hypothetical challenges, and responses) and to reflect on their significance to the argument's merits better than a purely text-based alternative. In an experiment, we found some empirical support for this hypothesis.

Introduction

Researchers in the field of intelligent tutoring systems have long been interested in developing systems that can help students learn through argument and debate or support the learning of argumentation skills [2, 7, 10].

One of the subtler aims of a law school education is to help students develop a "legal imagination." When law students hear of a proposed legal rule for guiding – or judging – behavior, can they imagine situations in which the proposed rule would lead to unintended results or conflicts with deeply held norms? Can students use these hypothetical examples to critique the proposed rule and can they respond to such critiques? Legal interpretation with hypotheticals is not only an important form of legal argument, it epitomizes something fundamental about the nature of legal reasoning: attorneys and judges reason *about* legal rules, not just *with* them. A proposed rule can be seen as a hypothesis, a tentative assumption made in order to draw out and test its normative, logical or empirical consequences. A hypothetical is an imagined situation that helps to test such a hypothesis; it is used to help draw out the various types of consequences of a proposed rule. U.S. Supreme Court Justices are famous for posing hypotheticals during oral arguments to evaluate proposed rules for deciding a case. As a step toward getting students to develop skills at reasoning with hypotheticals, the LARGO ("Legal ARgument Graph Observer") intelligent tutoring system supports

them as they study U. S. Supreme Court argument transcripts in which expert reasoning with hypotheticals unfolds. LARGO enables students to represent these arguments in simple graphic terms and then to reflect on the arguments' merits and significance.

Researchers aiming to develop systems that engage students in argument or improve their argumentation skills have been drawn to graphical representations for a number of reasons. From a cognitive perspective, graphical representations can reduce the students' cognitive load and reify important relationships. Thus, it is hypothesized, they facilitate reasoning about texts and the acquisition of interpretive skills [1, 5]. While the use of two simultaneous representations can increase cognitive load, the complementary strengths of a textual and graphical argument form can better guide students in their analysis. Second, intelligent tutoring systems can provide feedback on graphical argument representations while finessing the fact that natural language processing remains difficult. A student-made graph provides the system with information about their thinking that, even if it does not rise to the level of complete understanding, can be leveraged to provide intelligent help [8, 9].

Unlike earlier systems, LARGO's diagrammatic language is geared toward the specific form of argumentation that it supports, as opposed to a general argument representation, enabling LARGO to provide more task-specific targeted feedback. We conducted a controlled experiment to test the hypothesis that LARGO's graphical representations and feedback help students learn better than with the help of a purely text-based tool. This paper presents the results of this system evaluation.

1. Example and Model of Legal Interpretation with Hypotheticals

We illustrate the process of legal interpretation with hypotheticals with an extract from the oral argument in the case of Kathy Keeton v. Hustler Magazine, 465 U.S. 770 (1984). This case deals with one of the first technical legal concepts that law students encounter in the first year: personal jurisdiction, a court's power to require that a person or corporation appear in court and defend against a lawsuit. These cases often pit the principle that a state may redress wrongs committed within the state against the U.S. Constitutional principle of "due process" (minimum procedural safeguards against the arbitrary exercise of government power), especially when a court sitting in one state asserts power over a nonresident of that state. The plaintiff, Kathy Keeton, sued Hustler Magazine, an Ohio corporation with its principle place of business in California, in U.S. District Court in New Hampshire. She claimed that Hustler had libeled her in five articles. She was not a resident of New Hampshire and had almost no ties there. Hustler's contacts with New Hampshire included the monthly sale of up to 15,000 magazine issues there. At the time, New Hampshire was the only state in which Ms. Keeton was not barred under a state statute of limitations from making her claim.

In U. S. Supreme Court oral arguments, each side gets one half hour to address the issue before the Court. The extract shown in Figure 1 illustrates key argument moves used during these sessions, modeled as in [4]. The left column labels the different argument elements, such as proposed tests, hypotheticals, and ways of responding to hypotheticals, while the right contains the actual argument text. "Q:" indicates a Justice's question. Mr. Grutman represents Ms. Keeton. He begins by proposing a rule-like test for deciding the problem in a manner favorable to his client (line 14). Such proposals often include supportive reasons, such as that the proposed test explains past case decisions or is consistent with, or best reconciles, principles and policies

underlying the law. Justices may respond by posing a hypothetical (lines 55, 57, 59), which may simply be a query about the test's meaning (line 55 & 57), or may underscore the test's overly broad scope (line 59). The advocate has to rebut or otherwise reply to the challenge to maintain his argument's credibility. He may attempt to justify his proposed test by arguing that the supposedly disanalogous counterexample (i.e., the hypothetical) is really analogous to the current fact situation (cfs), in effect disputing that the proposed rule should yield a different result when applied to the counterexample than when applied to the cfs (as in lines 56 & 58). Or, he may distinguish the hypothetical from the cfs (as in lines 64 & 66).

→ **Proposed test of Mr. Grutman for Plaintiff Keeton**	14. GRUTMAN: The synthesis of those cases holds that where you have purposeful conduct by a defendant directed at the forum in question and out of which conduct the cause of action arises or is generated that satisfies the formula of those minimum contacts which substantial justice and reasonable fair play make it suitable that a defendant should be hailed into that court and be amenable to suit in that jurisdiction.
← **J.'s hypo**	55. Q: Would it apply in Alaska?
→ **Response: analogize cfs/hypo**	56. GRUTMAN: It would apply, Mr. Justice Marshall, wherever the magazine was circulated. It would apply in Honolulu if the publication were circulated there. It would apply theoretically and, I think, correctly wherever the magazine was circulated, however many copies were circulated
← **J.'s hypo**	57. Q: Just to clarify the point, that would be even if the plaintiff was totally unknown in the jurisdiction before the magazine was circulated?
→ **Response: analogize cfs/hypo**	58. GRUTMAN: I think that is correct, Mr. Stevens, so long as Alaska or Hawaii adheres, I believe, to the uniform and universal determination that the tort of libel is perpetrated wherever a defamatory falsehood is circulated. Wherever a third person reads about it, there is that harm
← **J.'s hypo**	59. Q: What if the publisher had no intention of ever selling any magazines in New Hampshire
	60. GRUTMAN: A very different case, Mr. Justice White.
→ **Response: distinguish cfs/hypo**	64, 66. GRUTMAN: … because in that case you could not say, as you do here, that you have purposeful conduct. There you have to look for other -- I think your phrase is affiliating circumstances, other connections, judicially cognizable ties --

Figure 1. Examples of interpretive reasoning with hypotheticals in oral argument in Keeton v. Hustler.

2. The LARGO Intelligent Tutoring System

From the viewpoint of legal pedagogy, oral argument examples like that above are worth studying, but they are challenging materials to beginning law students. A program that engages students in reflecting upon such expert examples could help; it could bring the general argumentation principles to the forefront and at the same time require that students be active learners. LARGO allows law students to graphically represent the dialectical pattern of hypothetical reasoning. Figure 2 shows an example graph based upon the Keeton case. This graph was prepared by a naïve user for the purpose of illustration (since Keeton was part of the post-test for our study the subjects did not use LARGO with this case). The left side of the screen shows parts of the oral argument transcript. Below that is an advice button and a palette of the basic graphical elements for markup. These elements include nodes representing proposed tests, hypotheticals, the cfs, and relations (like modification, distinction and analogy). Graphs are constructed by dragging these elements into the workspace and filling in appropriate text. Students can also link graph elements to parts of the transcript using a highlighting feature (e.g., the yellow node in Fig. 2 is linked to line 58 of the transcript).

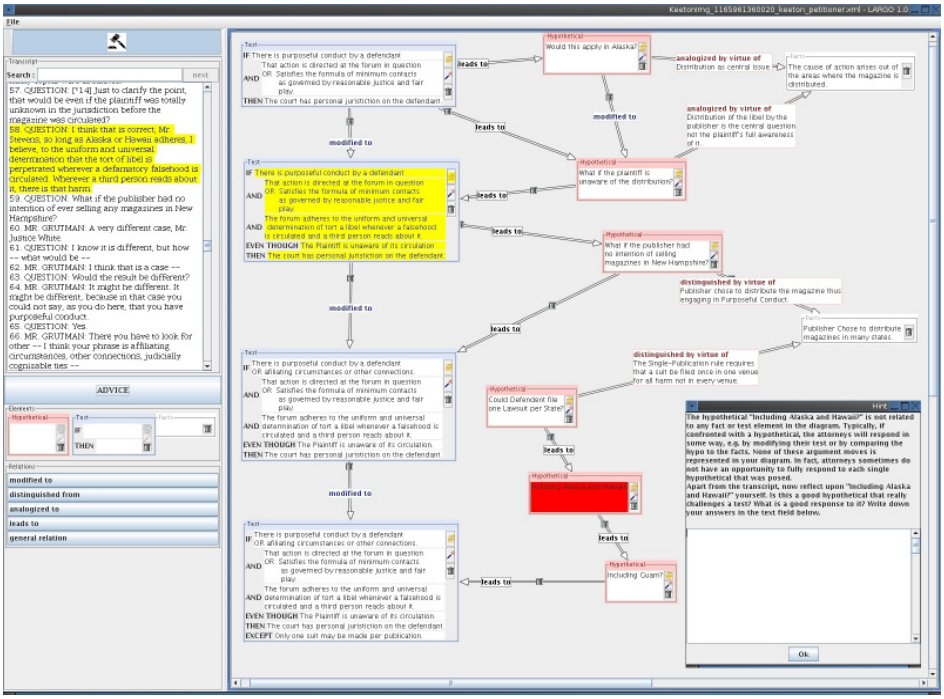

Figure 2. LARGO Representation of Keeton Case Oral Argument

When a student clicks on the Advice button, LARGO provides advice on improving their diagram. It detects three categories of diagram characteristics (a characteristic can either be a potential diagram weakness [9] or an opportunity for reflection). *Structural* characteristics involve portions of the graph where the relations among elements either do not correspond to the model of interpretive reasoning illustrated in section 1 (weaknesses) or correspond to an interesting or unusual constellation (opportunities for reflection). For example, hypotheticals are typically related directly to a test or other advocate response during oral argument. Novices often fail to represent these relations with diagram links. LARGO is capable of detecting this kind of diagram characteristic, and considers it a weakness. In Figure 2 LARGO provides a structural hint in the dialog box at the bottom right, pointing out that the hypothetical "What about Alaska and Hawaii" is not connected to any test element. *Context* weaknesses involve situations where the student's graph does not contain elements corresponding to key passages in the argument transcript that have been marked up in advance by a law professor. For example, the student may have missed a proposed test, or may have represented a test as a hypothetical. Finally, *content* weaknesses are poor formulations of proposed tests identified in the transcript.

Internally, LARGO relies on two analysis mechanisms: First, a graph-grammar formalism is used to detect context and structural weaknesses or opportunities for reflection. Second, a collaborative filtering technique is used to remediate content weaknesses. When students, after reading an attorney's proposed test in the transcript, enter a formulation of that test into their diagram, they are prompted to rate their formulation of that test against others produced by their peers. This information enables LARGO to derive a ranking of the formulations of a given test by all students [9].

Given the ill-defined nature the legal domain [6], one cannot always be certain that a diagnosed graph weakness represents an inaccurate rendition of the transcript, or how it should be "repaired". It may be that the particular line of argument is unusual (it is difficult to foresee all such possibilities). Or the Justices may have abandoned a line of questioning before a standard argument pattern could be completed. Therefore, LARGO's feedback is typically couched as invitations to reflect or as self-explanation prompts. These types of prompts have proven effective as a metacognitive strategy also in ill-defined domains [12]. For example the hint in Figure 2 prompts the student to think about the fact that one of the hypotheticals, highlighted red, is unconnected to any test or fact. If that was indeed the case in the transcript, then the diagram should reflect that (and thus is fine as it is), but if not the advice may lead the student to repair it. In either case it seems useful to invite the student to reflect on the role of this hypothetical.

3. Experimental Procedure

We conducted an experiment comparing LARGO's graphical representations and advice with a text-based alternative which simulates the process of examining the text with a notepad by allowing students to highlight portions of the text and enter notes in a text pane (without feedback). Our hypothesis was that the graphical format and advice would help students better identify and understand the argument components. The experiment was conducted in concert with the first-year Legal Process course at the University of Pittsburgh and with the professors' permission. The cases examined in the study centered on personal jurisdiction which was part of their coursework. We invited students to volunteer for the study, they received $80 for their participation. Students were assigned randomly to the conditions, balanced in terms of LSAT (Law School Admissions Test) scores. 38 students began the study, 28 completed it.

The experiment involved four sessions of 2 hours each over a four-week period. The first was a pre-test and an introduction to the software. In the second and third sessions, the students worked with extracts of the oral arguments from two personal jurisdiction cases. In the Experimental condition, students represented them graphically using LARGO with the help of the feedback mechanisms. In the Control condition, students were instructed to highlight relevant passages and take notes using the notepad. Session four consisted of the post-test. The pre- and post-tests, designed to assess students' argument and hypothetical reasoning skills, comprised five types of multiple-choice questions: 1) Legal argument-related questions of a type considered for inclusion in the LSAT [11]; 2) General questions about the use of hypotheticals in legal argument; 3) Questions that explored the use of hypotheticals for argument in a non-legal, intuitive domain about the policies of a tennis club; 4) Near-transfer argumentation questions about tests, hypotheticals, and responses in a new personal jurisdiction case (Keeton); 5) Far-Transfer argumentation questions drawn from a legal domain (copyright law) with which students were not likely to be familiar. The first three types appeared on both pre- and post-test; the last two only on the post-test.

4. Results and Discussion

We first computed a single overall pre-test and post-test score for each subject, which included all survey items with multiple choice answers. We also computed subscores

for each of the five specific question types described above. No significant difference on overall or subscores existed between the two conditions on the pre-test. The post-test scores for the Experimental subjects were consistently higher than the Control subjects' scores both with respect to overall scores and subscores. However, these differences were not statistically significant ($t(1,26)=.92$, $p>.1$, for overall scores). For some question types, the effect size between conditions would have been considerable if the results had reached the level of significance ($t(1,26)=1.22$, Cohen's $d=.60$, $p>.1$, near transfer problem; $t(1,26)=1.25$, $d=.45$, $p>.1$, "Tennis Club" post-test questions).

We then divided up students by LSAT score, creating a "Low" group containing 10 students, a "Med" group with 9, and a "High" group with 8 (the group sizes vary slightly to ensure that all students with the same score are in the same group; all students in the Med group had the same LSAT score). One student did not have an LSAT score and was not included. The results of these three groups differed considerably ($F(2,25)=4.15$, $p<.05$, for the overall score, similar results for most subscores). The students in the Low group (average post-test score .54) scored significantly lower than those in the High group who had an average of .63 ($p<.01$), consistent with the predictive value claimed for the LSAT scores. The Med group was in between the two (average .55). In a subsequent analysis we found that for the students in the Low group, there was a condition effect for the near-transfer questions (Keeton case). We hypothesized that lower aptitude Experimental subjects would do better than the lower aptitude Control subjects. A 1-sided t-test showed an effect size of 1.99 ($p<.05$). There were no pre-test differences between these groups ($p>.8$).

In order to determine whether LARGO helped students understand some parts of the argument model better than others, we classified the questions in the pre- and post-tests in terms of which aspect of the argument model they relate to most: tests, hypotheticals, relations between the two, responses to hypotheticals, or none (general questions). This post-hoc analysis revealed another effect: students in the Low and Med groups benefited from LARGO and did better on post-test questions that ask them to evaluate a hypothetical with respect to a given test. For the Low group and the combined Low+Med group, the difference was significant (Low+Med: $t(1,17)=2.73$, $d=1.00$, $p<.05$, 1-sided), but not for the whole group (Low+Med+High). The differences between conditions for other LSAT groups and test items were not significant. Below is an example of a "hypothetical evaluation with respect to test" question where the Experimental subjects outperformed the Control subjects:

> Assume that Mr. Grutman's proposed test is as follows: If … defendant has engaged in purposeful conduct …, and … satisfies the minimum contacts … (see Figure 1, line 14).
>
> The following hypothetical was posed in the oral argument. It is followed by four explanations why the hypothetical is or is not problematic for Mr. Grutman's proposed test. Please check ALL of the explanations that are plausible.
>
> "… if the plaintiff was totally unknown in the jurisdiction before the magazine was circulated?" (see Figure 1, line 57)
>
> - o The hypothetical is problematic for Mr. Grutman's proposed test. The decision rule applies by its terms, but arguably the publisher should not be subject to personal jurisdiction in the state under those circumstances.
> - o The hypothetical is not problematic for Mr. Grutman's proposed test. The decision rule applies by its terms, and the publisher should be subject to personal jurisdiction in the state under those circumstances.
> - o …
> - o …

These results support our research hypothesis (although perhaps fall somewhat short of decisively confirming it). For the Low group, the use of LARGO with its graphical argument representation and feedback in the form of tailored self-explanation prompts led to significantly better learning of argumentation skills than the use of traditional note-taking techniques, as measured in a near transfer problem which involved argumentation questions about a novel case in the same legal domain as those studied in the experiment. For the far transfer problem (a novel case in different legal domain) this effect was not found. We are also intrigued by the finding the Low+Med Experimental subjects apparently learned more about evaluating hypotheticals with respect to tests (not restricted to near transfer questions) than their Control counterparts. As the example illustrates, this skill is central to what LARGO is designed to teach: The relationship between tests and hypotheticals is the essence of oral argument.

One important question is why a significant difference was found on this particular question type and not on the other argumentation-model-related items (tests, hypotheticals, responses to hypotheticals). One possible explanation is that LARGO's graphical language distills the essence of the oral argument visually, explicitly identifying the relations between tests and hypotheticals (cf. Figure 2). Our data suggests that less skilled students benefited from creating and reflecting on these diagrams (with the help of LARGO's feedback), whereas more skilled students may have been able to understand the complex relations without aid. One can argue that for the other argumentation-model-related items, the specific graph structure or advice features that LARGO employs are not sufficient to differentiate it from purely text-based annotation tools. The student's ability to formulate a good test might not be supported to a great extent by a graphical representation format or prompts LARGO offers. However, as the near-transfer effect for the Low group shows, the less skilled students do benefit from LARGO also on a general level.

In summary, our analysis of the study results so far shows that the LARGO ITS is a valuable tool for those learners who do not (yet) have the ability to learn argumentation skills from independent study of argument transcripts. This group seems to benefit from the scaffold that the diagrams and the feedback offer. For the more advanced/skilled students, LARGO did not prove to be significantly better (but also not worse) than traditional learning resources such as a notepad and a highlighter. Yet, an analysis of the log files indicates that this group too appreciated LARGO's feedback and advice functions. On average (across all sessions of the study and all students in the Experimental condition), the advice button was pressed 10.1 times per transcript (i.e., per hour). All 3 groups frequently requested advice: Low, 12.3; Med 6.2; High 17.9. In 75% of these cases, students selected one of the three short hint titles that LARGO presented in response to their hint request, and read through the detailed feedback related to the selected hint title. The use of the advice did not decrease over time. In the later sessions, the average number of help requests was even higher than in the earlier sessions (12.2 and 8.6 in the last two transcripts vs. 7.3 and 9.8 in the first two), evidence that the students must have considered the advice to be valuable.

5. Conclusions and Outlook

A study carried out in a first-year law school course provides support for the hypothesis that a diagrammatic language, combined with feedback that points out weaknesses and opportunities for reflection in students' argument diagrams, helps students learn to

apply a general model of hypothetical reasoning, as they study transcripts of arguments made by highly-skilled experts. For lower-aptitude students, use of LARGO's diagramming and feedback functions was more effective than traditional note-taking techniques. Specifically, within this group, those who used LARGO learned better to analyze new argument transcripts in the same area of the law, even when they studied the new transcript without the use of LARGO. They also learned better to reason about how a hypothetical might relate to a proposed test, a key element of hypothetical reasoning. The study thus demonstrates the benefits and potential of graphical representations in intelligent tutoring systems for argumentation, especially with lower-tier students. In an earlier study [3], we found benefits of (non-dynamic) self-explanation prompts based on the same general model of hypothetical reasoning on which LARGO is based. This study was carried out in the context of a program to help disadvantaged students prepare for law school. Taken together, these studies suggest that LARGO would be a valuable addition to regular law-school curricula and to preparatory programs for disadvantaged students.

At present we are continuing the analysis of the study data. We will conduct further analyses of students' log files, argument graphs, and text notes. We are particularly interested in determining whether the Experimental subjects were better able to find and relate key tests and hypotheticals than their Control counterparts. We will also determine how the variability of students' argument graphs speaks to the ill-defined nature of the domain [6]. Furthermore, we will investigate which of LARGO's hints were most helpful. LARGO represents one of the first full-scale implementations of the collaborative filtering idea in an educational application. We will verify whether the filtering did, in fact, accurately assess students' formulations of the tests they identified in the argument transcripts.

References

[1] Ainsworth, S. (1999). The functions of multiple representations. *Computers and Education 33,* 131-152.

[2] Aleven, V. (2006). An intelligent learning environment for case-based argumentation. *Technology, Instruction, Cognition, and Learning*.

[3] Aleven V., Ashley, K.D., & Lynch, C., (2005). Helping Law Students to Understand US Supreme Court Oral Arguments: A Planned Experiment. Proc. of 10th Int. Conf. on AI and Law. Bologna Italy.

[4] Ashley, K. (2007). Interpretive Reasoning with Hypothetical Cases. In Proc. 20th Int'l FLAIRS Conference, Special Track on Case-Based Reasoning, Key West, May.

[5] Larkin, J. & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science, 11,* 65-99.

[6] Lynch, C., Ashley, K, Aleven, V., & Pinkwart, N. (2006). Defining Ill-Defined Domains; A literature survey. In V. Aleven, K. Ashley, C. Lynch, & N. Pinkwart (Eds.), Proc. Workshop on Intelligent Tutoring Systems for Ill-Defined Domains at the 8th Int'l Conf. on Intelligent Tutoring Systems, 1-10.

[7] Muntjewerff, A. J., & Breuker, J. A. (2001). Evaluating PROSA, a system to train solving legal cases. In J. D. Moore et al. (Eds.), *Proc. of AIED 2001* (pp. 278-285). Amsterdam: IOS Press.

[8] Paolucci, M., Suthers, D., & Weiner, A. (1996). Automated advice-giving strategies for scientific inquiry. In C. Frasson, G. Gauthier, & A. Lesgold (Eds.), Proc. ITS '96 (pp. 372-381). Berlin: Springer.

[9] Pinkwart, N., Aleven, V., Ashley, K., & Lynch, C. (2006). Toward Legal Argument Instruction with Graph Grammars and Collaborative Filtering Techniques. In M. Ikeda, K. Ashley, & T. W. Chan (Eds.), Proceedings of ITS 2006 (p. 227-236). Berlin: Springer.

[10] Ravenscroft, A. & Pilkington, R.M. (2000) Investigation by Design: developing dialogue models to support reasoning and conceptual change. *Int. Journal of AI in Education, 11*(1), 273-298.

[11] Reese, L. & Cotter, R. (1994) A Compendium of LSAT and LSAC-Sponsored Item Types 1948-1994. Law School Admission Council Research Rept. 94-01.

[12] Schworm, S., & Renkl, A. (2002). Learning by solved example problems: Instructional explanations reduce self explanation activity. In Proc. Ann. Conf. of the Cogn. Sci. Society (pp. 816-821). Erlbaum.